Quantitative Studies of Place and Spatial Regression Analysis: A Study of the Legacy of Slavery in the U.S. South

Contributors: Heather A. O’Connell & Robert L. Reece
Pub. Date: 2018
Access Date: January 12, 2018
Academic Level: Postgraduate
Publishing Company: SAGE Publications Ltd
City: London
Online ISBN: 9781526441164
DOI: http://dx.doi.org/10.4135/9781526441164

Abstract

Space and place are important concepts when studying how society functions. Spatial regression analysis techniques provide one important means for incorporating space into quantitative analyses. We highlight this methodological approach by detailing the steps that we took in a recent study of the legacy of slavery. Our case study provides guidance on when and how to use spatial statistics like the Moran's I statistic, Lagrange Multiplier diagnostic tests, and spatial regression models. Our previous research on the legacy of slavery offers a prime example of how accounting for space may be important even if it is not the focus. The goal of our study was to better understand the legacy of slavery, but to estimate the focal relationship using regression analysis we needed to account for any unobserved processes related to the spatial position of our units of analysis. Sociological scholars in all areas need to consider how space may be affecting their statistical results. Furthermore, we argue that closer attention to spatial position through other methodological approaches–both quantitative and qualitative–can provide greater clarity on the social processes that we aim to understand.

Learning Outcomes

By the end of this case, students should be able to

- Identify when spatial regression analysis techniques are necessary
- Understand the difference between a spatial error and spatial lag regression model
- Apply spatial perspectives to their own research questions

Project Overview and Context

In this methods case study, we are going to discuss the steps that we took in a project that focuses on the relationship between history and contemporary racial inequality (Reece & O'Connell, 2016). We were particularly interested in whether and why counties in the U.S. South with higher historical concentrations of slaves continue to have greater disparities in Black-White public school attendance. Our focus on places (i.e., counties) leads to some important methodological issues related to space that have broad applicability in Sociology.

Before discussing the methods, we provide a little background on the project. Our main goal was to better understand the contemporary legacy of slavery. To do that we needed to pay attention to additional processes that could help explain the initial relationship between history and the contemporary. We focused on three core issues discussed in previous research: private schools, racial threat, and sub-regional variation.

1. Private schools, including "segregationist academies" that were explicitly developed to maintain White educational advantages, have been an important component of school segregation in the U.S. South (Clotfelter, 2004; Porter, Howell, & Hempel, 2014) and may be a part of the contemporary legacy of ideologies attached to slavery.
2. White responses to larger Black populations (i.e., racial threat and "white flight") are also closely connected to Black-White school segregation, especially when explaining why segregation levels differ across places (e.g., Blalock, 1967; Renzulli & Evans, 2005).
3. Finally, previous research suggests that Black population concentration matters more in some regions than others, and that there is a strong distinction between the Deep South and the other southern states (e.g., Tolnay, Beck, & Massey, 1989).

Our primary hypothesis was that places with stronger attachments to slavery (i.e., larger historical concentrations of slaves) would have a larger Black-White gap in public school attendance—a larger percentage of Black school-aged children would be enrolled in public schools than White school-aged children. However, we also expected that this initial relationship would be partially explained by the number of private schools in a county, and White responses to the relative size of the Black population.

There are a number of methodological challenges involved in testing our hypotheses about the connections among history, racial inequality, and place. In this case study, we focus on the challenges related to space. Space—or the position of places, people, or objects in relation to other entities—can be factored into a study in a variety of ways. Our methods cover a specific approach to incorporating space, namely through spatial regression analysis techniques, but we also discuss the broader concepts of space and place and their relevance to all sociological studies.

Research Design: What Additional Steps Do I Need to Cover When Focusing on Places?

Selecting a Unit of Analysis

The first question that we needed to address in our study of places is related to the appropriate unit of analysis. How should we define "place"? Our response to this question was shaped by our understanding of which unit best captures the processes that we were studying (i.e., public school attendance and the contemporary legacy of slavery), previous research, and data restrictions.

One intuitive geographic unit when analyzing dimensions of school segregation is the school district, but we argue that this unit is too small for capturing the broader dynamics related to the legacy of slavery and the context of private school availability. Cities may have offered a good alternative because they covered some of that broader context, but focusing only on cities would exclude the rural areas of the South where we might expect many of our hypothesized dynamics to unfold.

Ultimately, we selected counties as our unit of analysis. Counties had been used in previous research on the legacy of slavery and have a number of other benefits, including the availability of a wide range of data and strong connections to southern school systems. However, there are also some limitations.

A primary limitation is that the boundaries defining a county can change over time—sometimes dramatically—which may raise some concerns regarding the social meaning of counties and other geographies defined by the Census (e.g., census tracts). Changing boundaries suggest some level of arbitrariness instead of a durable, socially defined place. Despite its limitations, the county is the best unit of analysis for our study because they most closely represent the social spaces that are the focus of our analysis.

Selecting the most appropriate available unit of analysis is an important first step in any analysis of place and requires careful consideration of the pros and cons. No pre-defined geographic unit will ever be perfect for capturing sociological processes, but they are often the best that we have available, especially when conducting large-scale quantitative research.

Variable Construction: Connecting History to the Present

The construction of most of our variables was straightforward and has little connection to the spatial challenges that are our focus in this case study. However, we list the dependent and key independent variables below:

- Dependent variable: ratio of the proportion of Black students to the proportion of White students enrolled in public schools
- The number of slaves relative to the total population, 1860
- The number of private schools per 100 students in a county
- An estimate of racial threat based on the ratio of the Black relative to the Black plus White populations
- A binary variable indicating whether a county is located in a state that was part of the original Confederacy

All of our variables, except for the slavery variable, are from the 2006-2010 American Community Survey (ACS) county estimates or another contemporary source. However, it was important for our analysis that the historical slave and total population variables from 1860 be close reflections of the contemporary county boundaries. The changes in county boundaries that can occur over time (also mentioned above) make accomplishing this complicated.

We used a tool developed by Adam Slez, O'Connell, and Curtis (2017) and a careful incorporation of the spatial position of a county to get the best estimate of the historical slave concentration of a contemporary county. The tool allowed us to identify quickly all of the counties that experienced a significant boundary change by comparing the county boundaries observed in 1860 and 2000.

Once we had that list, we could cross check the estimates of slave concentration for those counties using maps for 1860 and 2000. Knowing the concentration of slaves in neighboring counties often helped us adjust our estimates because places that are close to one another are often more similar than far places. (This idea—sometimes referred to as the first "law" of geography—will be important later on, too.)

Data Analysis

We selected regression as our method of data analysis because it allows us to isolate the relationship between historical slave concentration and contemporary disparities in public school attendance. However, standard ordinary least squares (OLS) regression analysis assumes that each observation is independent from the others, and our county observations *could* violate that assumption if there is spatial dependence in our residuals.

As we mentioned above, near things are more similar to one another than distant things. Spatial position is ignored in standard OLS regression models, but similarity related to space can still affect the model results. If we do not account for residual spatial dependence, then our estimates could be biased.

Understanding and Identifying Important Spatial Processes

Two neighboring counties may share more in common with one another than we originally specified in our statistical models. Unobserved similarity would result in correlated residuals—a violation of the OLS assumption of independence—and because that correlation is due to the spatial proximity of the observations we call it "spatially correlated residuals" or "residual spatial autocorrelation."

Further complicating the issue of spatial position is the fact that the similarity of two near places could result from one of two types of spatial processes, namely a spatial error or a spatial lag process. A spatial error suggests a misspecification related to either your independent variables or your unit of analysis, but a spatial lag suggests a social process related to your dependent variable.

A spatial error process could be due to the omission of a variable or it could be a signal that the unit of analysis is not a good fit. For example, knowing that there is a political organization with a jurisdiction that spans three counties may help explain the greater than expected similarity of those counties. Furthermore, that similarity could suggest that the social processes explaining the value of your dependent variable are not restricted to the county unit and are better represented by the larger jurisdiction.

A spatial lag process is slightly different in the context of explaining residual spatial autocorrelation because it emphasizes the diffusion or spread of your dependent variable from place to place. Spatial lag processes may be particularly relevant when you are studying attitudes, for example, because the attitudes expressed in one place could be affected easily through direct social interaction with residents in neighboring places. Another common example of diffusion can be found in research on the spread of disease. When explaining the prevalence of a contagious disease, accounting for the prevalence in neighboring places is critical. The same is true when studying “contagious” social processes.

What to Do

There are three basic steps that need to be taken when you think that spatial processes may affect your regression analysis:

1. Estimate the magnitude and significance of correlation among your model residuals using a Moran’s *I* test statistic
2. Distinguish between the two alternative approaches (i.e., spatial error and spatial lag) for dealing with residual spatial autocorrelation using a Lagrange Multiplier test
3. Conduct regression analysis using the appropriate model

Estimating and Interpreting the Moran’s I Statistic

First, we ran the standard OLS regression model and estimated the Moran’s *I* statistic in R software. R is known for its flexibility and the availability of new programs developed by users, but spatial modeling techniques are available through other software platforms, too (e.g., Stata, ArcGIS).

The Moran’s *I* statistic requires what is referred to as a “spatial weights matrix.” This matrix identifies which places count as a neighbor. Your matrix can be defined in a number of ways. The first relies on distance. Any place within a specified distance would be considered a “neighbor” of that county. The second selects a predetermined number of the nearest neighbors—this is the “k nearest neighbors” approach. The third approach builds a “contiguity” matrix, which is based on whether the boundaries of two places are touching.

We opted to use a contiguity matrix in our analysis. We had no reason to expect that a certain distance would be especially relevant to the processes that we were studying, so that ruled out the first distance approach. We could have used the “k nearest neighbors” approach, but we chose not to because you have less control over the actual proximity of the neighbors that get selected when using that approach.

The Moran’s *I* statistic for our model residuals was 0.07, which is a modest level of spatial autocorrelation (i.e., less than 0.10), but it was still highly statistically significant (*p* < .001). This suggests that there are unobserved spatial processes that may be affecting our coefficient estimates. If we had found a non-significant Moran’s *I* statistic, then we would have continued with a standard OLS model and skipped the remaining steps.

Distinguishing Between Spatial Error and Spatial Lag

Second, we considered our options for treating that residual spatial autocorrelation. Is it resulting from processes related to an error or a lag specification? The range of factors that could contribute to a spatial error pattern is much wider than the processes tied to a spatial lag, so we started by addressing theoretically whether we thought a diffusion process could be involved.

Our dependent variable was the ratio of Black relative to White public school attendance in a county. It is possible that the extent of this disparity in one county could contribute to the value observed in a neighboring county, at least indirectly by affecting norms related to private school attendance. However, our model already includes some variables that we think might capture those kinds of norms (i.e., historical slave concentration). It is more likely that the small, remaining spatial autocorrelation is driven by error processes.

We used the Lagrange Multiplier diagnostics test to distinguish empirically between the error and lag options. Sample code is provided below:

```
lm.LMtests(ols_model, weights_list, test=c(“LMerr”,”RLMerr”,”LMlag”,”RLMlag”))
```

The bolded text indicates where you would insert the name of the object that contains the results from your OLS model (e.g., ols_model) and the object that contains the list version of your spatial weights file (e.g., weights_list).

This and related functions are available through the R package “spdep.” Helpful guides providing more complete information on how to conduct this kind of analysis are available in print and online. Some suggestions are offered in the “Further Readings” and “Web Resources” sections below.

The Lagrange Multiplier diagnostics test provides five statistics, but we will focus only on the first four because they are the most relevant. The LMerr statistic is a test for error dependence. The LMlag statistic tests for whether you should include a spatially lagged dependent variable. The second two statistics—RLMerr and RLMlag—are the robust versions of the first two tests and are only used when the LMerr and the LMlag tests are inconclusive.

When we used the Lagrange Multiplier diagnostics test in our analysis, we found that the tests for LMerr and LMlag were both statistically significant. That means both processes—an error and a lag—could be involved. However, we want to know which one is the dominant process, so we turn to the robust tests because those statistics isolate each process from the other (e.g., the RLMerr tests for an error process net of a lag process).

In our study, the RLMerr coefficient was significant ($p < .01$) and the RLMlag was not. This gave us a clear indication that we should use a spatial error regression model. If your results are less clear than our own—say, both robust statistics are significant—then you should choose the one with the highest level of significance (but probably estimate your models using both approaches as a sensitivity check).

Conducting the Regression Analysis

The third and final step is the easiest: we estimated our regression models using a spatial error specification. The interpretation of the model is nearly identical to that of a standard OLS model. Spatial error models include an additional coefficient to represent the unobserved spatial process, but it has little substantive meaning. If we had needed to use a spatial lag model, then we would have more to say about the spatial coefficient because it would be tied to the idea of spatial diffusion.

Do I Need to be concerned About Spatial Processes?

You should be concerned about residual spatial autocorrelation any time that you are using regression analysis and your units of analysis are close to one another spatially. For example, if you are studying all of the neighborhoods in a city, then the relationships that you are trying to estimate may be affected by unobserved similarities across neighboring areas.

We focused on the technical issues related to spatial regression analysis, but residual spatial autocorrelation is only one part of how spatial processes can impact your research. You should also consider the positive role that space can play—how can thinking about spatial position improve my understanding of this social process?

There are many ways to improve our research by thinking about space. Spatially lagged variables are a common tool used in Sociology because of their connection to diffusion processes. David Cunningham and Benjamin Phillips (2007), for example, used a spatial lag approach to show how the presence of Klu Klux Klan (KKK) activity in neighboring counties increases the odds of a KKK presence in other counties in North Carolina.

Even more detailed than a spatial lag approach, Derek Alderman (2000) used information on the spatial positioning of specific roads to enhance his study of how Martin Luther King Jr. has been commemorated in the U.S. South. By distinguishing between renaming a five mile road that stretches through downtown and a half mile road in an isolated residential area, his work emphasizes the importance of *where* something is done in addition to *whether* something is done. He argues that the spatial prominence or invisibility of a commemorative symbol gives us greater understanding of the social processes underlying the commemoration of historical figures like Martin Luther King Jr.

Conclusions

Spatial perspectives are always helpful—everything that we study is unfolding in place and within a spatial structure—but spatial data analysis techniques are not always necessary. In this case study, we provide guidance for understanding and identifying spatial processes that may affect regression analyses.

Studies of place require special attention to defining appropriate boundaries for capturing the underlying social processes. Places are also inherently connected to one another through a spatial network, which may affect your results whether you had intended to study spatial processes or not. Accounting for space is necessary even when space is not a focus.

Several tools have been developed to help identify and address spatial processes in quantitative research. The Moran's *I* statistic is a helpful first step for identifying the extent of spatial dependence. When faced with significant residual spatial autocorrelation, we need to then consider the type of spatial process that is driving that dependence—spatial error or spatial lag.

A spatial error approach to residual spatial autocorrelation treats the spatial dependence as a "nuisance" in some sense because it does not come with a specific interpretation. A spatial lag, in contrast, focuses on diffusion processes. The implication of a spatial lag approach is that the value of the dependent variable in one place is somehow affecting the value in neighboring places.

We should first distinguish between the spatial error and spatial lag processes theoretically—which makes most sense for the process that you are studying?—but then we can use the Lagrange Multiplier test to distinguish them empirically. This test provides the final guidance before deciding what type of model best suits your data.

While the spatial methods that we discussed here are more specialized, the concept of space has broad applicability. Regardless of the method that you are using—quantitative or qualitative—you should consider the role of spatial positioning. Geographic continuities and obstructions can affect social connectivity, and the spatial position of an object signals its broader social importance. Incorporating space and place into sociological studies is critical to advancing our understanding of society.

Exercises and Discussion Questions

1. What types of processes are associated with a spatial error and spatial lag?
2. The current case study only focuses on the spatial lag in the context of the dependent variable, but it can also be used to develop independent variables. Using your understanding of the processes related to a spatial lag, explain how you could use a spatially lagged *independent variable* to test a hypothesis about the relationship between public school attendance in a county and private school availability. What would your hypothesis be and how would you test it?
3. How would you define the spatial weights matrix for the project developed in question 2?
4. Besides concerns about how unobserved spatial processes might affect our regression results, what are some other ways that space could affect how you think about your own research?
5. How would you describe the similarities and differences between the spatial data analysis techniques discussed in this case study and multilevel modeling techniques that account for the clustering—and therefore greater than expected similarity—of observations?

Further Reading

Cliff, A. D., & **Ord, J. K.** (1973). *Spatial autocorrelation*. London, England: Pion Limited.

Cliff, A. D., & **Ord, J. K.** (1981). *Spatial processes: Models & applications*. London, England: Pion Limited.

Curtis, K. J., & **O'Connell, H. A.** (2016). Historical racial contexts and contemporary spatial differences in racial inequality. *Spatial Demography*, 5, 73–97. doi:http://dx.doi.org/10.1007/s40980-016-0020-x

Drukker, D. M., **Peng, H.**, **Prucha, I. R.**, & **Raciborski, R.** (2013). Creating and managing spatial-weighting matrices with spmat command. *The Stata Journal*, 13, 242–286.

Gieryn, T. F. (2000). A space for place in sociology. *Annual Review of Sociology*, 26, 463–496.

Logan, J. R. (2012). Making a place for space: Spatial thinking in social science. *Annual Review of Sociology*, 38, 507–524.

O'Connell, H. A. (2015). Where there's smoke: Cigarette use, social acceptability, and spatial approaches to hierarchical linear modeling. *Social Science & Medicine*, 140, 18–26.

Slez, A., **O'Connell, H. A.**, & **Curtis, K. J.** (2017). A note on the identification of common geographies. *Sociological Methods & Research*, 46, 288–299.

Voss, P. R., **Long, D. D.**, **Hammer, R. B.**, & **Friedman, S.** (2006). County child poverty rates in the US: A spatial regression approach. *Population Research and Policy Review*, 25, 369–391.

Ward, M. D., & **Gleditsch, K. S.** (2008). *Spatial regression models*. Los Angeles, CA: SAGE.

Web Resources

https://github.com/aslez

http://www.csiss.org/gispopsci/workshops/2011/PSU/readings/W15_Anselin2007.pdf

http://spatialdemography.org/software-and-code-spatial-analysis-in-r-part-2-performing-spatial-regression-modeling-in-r-with-acs-data/

References

Alderman, D. H. (2000). A street fit for a king: Naming places and commemoration in the American South. *Professional Geographer*, 52, 672–684.

Blalock, H. M. (1967). *Toward a theory of minority group relations*. New York, NY: Wiley.

Clotfelter, C. T. (2004). Private schools, segregation, and the southern states. *Peabody Journal of Education*, 79(2), 74–97.

Cunningham, D., & **Phillips, B. T.** (2007). Contexts for mobilization: Spatial settings and Klan Presence in North Carolina, 1964-1966. *American Journal of Sociology*, 113, 781–814.

Porter, J. R., **Howell, F. M.**, & **Hempel, L. M.** (2014). Old times are not forgotten: The institutionalization of segregationist academies in the American South. *Social Problems*, 61, 576–601.

Reece, R. L., & **O’Connell, H. A.** (2016). How the legacy of slavery and racial composition shape public school enrollment in the American South. *Sociology of Race and Ethnicity*, 2, 42–57.

Renzulli, L. A., & **Evans, L.** (2005). School choice, charter schools, and white flight. *Social Problems*, 52, 398–418.

Slez, A., **O’Connell, H. A.**, & **Curtis, K. J.** (2017). A note on the identification of common geographies. *Sociological Methods & Research*, 46, 288–299.

Tolnay, S. E., **Beck, E. M.**, & **Massey, J. L.** (1989). The power threat hypothesis and Black lynching: ‘Wither’ the evidence? *Social Forces*, 67, 634–640.